HyperTransport Technology: Optimized for DV & Audio

by The AMD Technology Evangelism Group
Advanced Micro Devices, Inc., One AMD Place, Sunnyvale, CA 94088

Introduction

Digital video (DV) and digital audio applications are putting strains on PC systems like never before. System performance is pivotal for compressing DV and audio into smaller, more manageable sizes, and for rendering effects. For consumer-level systems, the processor and accompanying system architecture must be able to decompress data pumped into the system, move it quickly between the processor and memory so the stream can be rendered with real-time special effects, and then recompress the stream into a suitable format for output. In professional systems, the movement of large *uncompressed* video within the system is the issue. If the system contains a bottleneck in any one of these areas, playback will be compromised and the compression of the data will cost extra precious time. Even the majority of the special-purpose hardware used for non-linear editing (NLE), where digitized video can be manipulated under software control, rely heavily on processors and system architectures.

The primary issue is not getting the DV and/or audio data into and out of the computer system; the issue is moving data around inside the system without compromising the integrity of the data stream. This is becoming more of a concern because processing power has continued to double every 18 months while the performance of the I/O bus architecture - the path along which the computer transfers data between system components - has lagged, doubling in performance approximately only every three years. Clearly, these two areas of technology are not being innovated at a comparable rate. If the system I/O bus cannot adequately move video frames from the memory to the processor to be modified, and then back to memory again so it can be displayed in a cohesive manner, there will be congestion. Creators typically solve this problem by fully rendering the video stream prior to viewing it, wasting large amounts of time.

Typically, the inherent scalability of the system means that the more processing power provided by a system, the better the overall performance. But as systems continue to be based on outdated I/O technology this will not be the case.

Inadequate Bus Architectures

Currently, systems use a limited bandwidth link comprised of a proprietary interface, between the Northbridge and the Southbridge. The Northbridge controls internal communications, and the Southbridge controls communications from hard drives and external importing devices. A limited I/O interface of this sort essentially creates a bottleneck as the chipset interconnect is forced to use a low bandwidth bus that must be used to share all external connections.

While these Northbridge and Southbridge connection technologies are capable of transferring data at throughputs from 266 Mbyte/s to about 1 Gbyte/s, and may contain smart electronics for better use of the bus, they still cannot handle well high-bandwidth data such as uncompressed DV, HD video or, more importantly, the compression and decompression of DV at the same time. Add concurrent audio encoding and other traffic and the buses all but grind to a halt.

HyperTransport Technology Solution
The core-logic architectures of today, particularly the interface between the Northbridge and the Southbridge, do not support isochronous data transfer in which packets are time-dependent and must be streamed in a way that ensures fluid playback. Lack of isochronous support is the leading cause for the inconsistent playback of DV and the stuttering of digital audio. Another issue is the lack of concurrency in today's interfaces. The inability to transfer data full-duplex is a major limiting factor in NLE, either using a software-based solution that taxes the processor and memory, or a hardware-based solution that relies on a dedicated capture card. When using a hardware solution the system must be able to handle very large files maintaining a continuous, uninterrupted flow of data both to and from the drive. Software solutions make this equation even tougher. HyperTransport I/O technology supports both isochronous and concurrent connections and has the capability to handle very large amounts of data at very high speeds.

Bandwidth, Bandwidth, Bandwidth
The motion picture and television industries are just beginning to work with DV. The growth of DTV and HDTV are placing an even-larger burden on the computers that are used for special effects and for compressing data for broadcast. With powerful NLE software tools available creators/producers can manipulate DV like never before - as long as the hardware is up to par.

I/O buses need to be capable of moving large amounts of DV. For full-screen, full-motion 525-line DV, the system must be capable of successfully moving 29.97 frame/s at 720 x 480 pixels in GBR 24-bit color, which equates to 31 Mbyte/s. A system built to handle HDTV (say, 1080i DV) must be capable of successfully moving 24 frame/s (standard being used by the motion picture industry) at 1920 by 1080 pixels for a total of 149 Mbyte/s. Theatrical quality resolutions (sic) are achieved by an interpolation being called "Up-RESing" where material is re-scaled by creating new pixels, comparing and averaging existing ones. Images are scaled to 2048 by 1151 pixels (at 24 frame/s) allowing images to fit into the larger area of 35-mm negative. As expected, this process requires much more processing power, with data rates equating to 170 Mbyte/s.

Of course, these bandwidths are for single streams of video: The need for multiple streams, especially when multiple video streams are used for creating transitions and when special effects are applied, places an even greater bandwidth burden on the system.

Compression Issues
Compression is the translation of source material, comprised of video, audio, or a combination of both, using a variety of computer algorithms to reduce the amount of data required to accurately represent the content, in a size that is manageable for storage and distribution.

While dedicated capture cards support hardware-based special effects and DV (DV) and MPEG compression codecs, they do not have all of the special effects and compression schemes required by the developer. These chores are left to software based plug-ins and applications typically bundled with the card while placing the majority of the processing load on the host processor. A typical scenario would be a DV capture card using a software codec to encode a video stream into an MPEG format, and sending the finished product to either the hard drive or an external device. An example of this is the DVStormSE capture card from Canopus.

Canopus DVStormSE Capture Card
The Canopus DVStormSE capture card provides real-time DV editing and boasts capabilities including the use of up to five independent video tracks, more than 20 title and graphic tracks, and 24 filters. It also employs Canopus's Scalable Technology Architecture to provide more features and increased performance as processor power increases. By relying on the processor and system architecture, the card can offer more real-time capabilities over time, ultimately increasing the number of real-time video tracks beyond five, and increasing the number of title and graphics tracks indefinitely and delivering more real-time creativity.

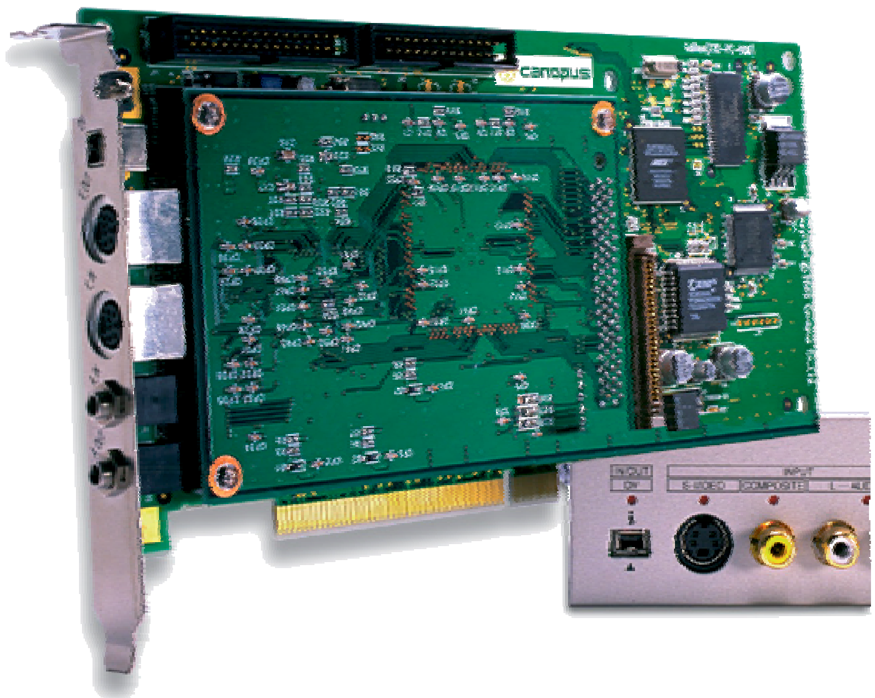

Fig. 1: Canopus DVStormSE Capture Card

I/O bandwidth continues to be an issue, whether DV relies on a hardware or software codec to handle compression and decompression chores. The data still needs to be moved within the system, and the bottleneck issue will continue to be a problem. This is especially true as DV sources become larger, and current bus technology continues to be unable to support high-speed transfers and concurrent transactions where data can be sent in both directions at the same time. DV is not the only area where this is important.

Reduced Audio Latency is Key
The high volumes, increased computing power, and reliability of personal computers has contributed to its outgrowth from the home studio into the professional studio as the platform used for the recording, editing, and mixing of high-fidelity audio content creation. In this environment, the real-time recording of multiple independent tracks, the mixing of special effects, and the real-time editing and playback of these tracks place tremendous strains on systems that must be capable of handling these streams without skips or glitches. While systems using older multi-drop, shared buses such as PCI can sustain today an aggregate of 20 Mbyte/s of raw audio data comprised of 48 tracks (24 in and 24 out simultaneously), each sampled with 24-

bit resolution at 96 kHz, the reality is these shared buses cannot always reliably handle these audio streams when other devices are sharing the bus.

Older multi-drop, shared I/O buses like PCI are half-duplex where data can only be sent in a single direction at a time while other devices wait their turn. Multi-drop buses, which can be connected to individual devices but divide the total available bandwidth between them, use an arbitration method that works on an interrupt basis with shared devices to guarantee bandwidth is distributed among them all. Compounding the issue further, many different arbitration schemes exist among architectures that can lead to inconsistencies between devices and dissimilar platforms. Also, inefficiency, bus turnaround, and overhead imposed on data sent across older shared buses like PCI greatly decreases the amount of overall bandwidth available to the bus and creates latencies, further encumbering fluid communication. This is not the optimal scenario for moving time critical audio streams without glitches. HyperTransport I/O technology, with its low latency and guaranteed high bandwidth attributes, is designed to solve these very issues.

NVIDIA's SoundStorm Audio Solution with Integrated Dolby Digital Encoder
Digital audio is experiencing the same fundamental issues as DV, albeit to a lesser extent. Audio innovation is making many new inroads as NVIDIA's nForce and nForce2 family of platform processors are a testament. The Media and Communications Processor (MCP/MCP-T) incorporates the NVIDIA nForce Audio Processing Unit (APU) which features the industry's first Dolby Digital Interactive Content Encoder (ICE.).. This technology can dynamically encode any multi-channel 2D or 3D audio source into Dolby Digital 5.1 in real-time and output it digitally. The nForce family of platform processors uses a high-speed HyperTransport I/O bus interface for internal communications, delivering unprecedented levels of PC performance.

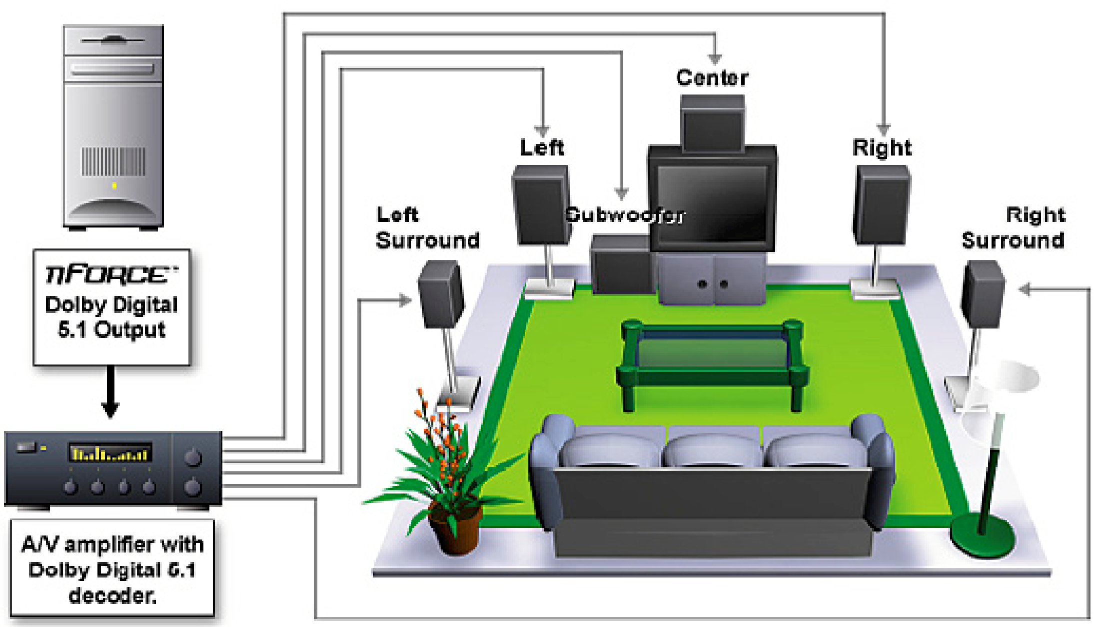

Fig. 2: SoundStorm Supports Six Speakers, Plus Dolby Digital Output via SPDIF

NVIDIA, finding that today's PC applications are increasingly complex with advanced 3D graphics, high-speed networking, streaming video, and cinematic 3D audio, found that no other

current technology allowed for the implementation of Dolby Digital 5.1 2D or 3D audio processing and broadband networking in the MCP. The APU integrates the Dolby Digital ICE into a programmable DSP with a fix-to-float format engine. This engine is used to take the output of the Global Processor and encode it into a Dolby Digital (AC-3*)* stream, allowing users to experience true theater-quality, multi-channel surround sound, *rendered in real-time,* on their Dolby Digital-equipped PCs, Mini-Disc players, and home theater systems.

Very High Bandwidth IN HyperTransport
HyperTransport links are capable of extremely fast signaling, designed to operate at clock speeds from 200 MHz up to 800 MHz. HyperTransport links use double data rate (DDR) technology and transfer two bits of data per clock cycle - an effective transfer rate of up to 1600 Mbyte/s in each direction. Since transfers can be full duplex an aggregate transfer rate of 6.4 Gbyte/s in a 16-bit HyperTransport I/O link and 12.8 Gbyte/s in a 32-bit link can be achieved.

Basically, the HyperTransport I/O link uses two point-to-point unidirectional links instead of the single-ended signaling employed by parallel buses like PCI. These two wires employ a signaling technique that reads data as the difference between the two signals, allowing operation at very high clock rates without suffering from issues possible on parallel buses: Bouncing, interference, and cross-talk from adjacent signals, which can potentially disrupt audio and video streams. The HyperTransport link is also a “packetized” bus, which means addresses, data, and commands are sent along the same wires allowing designers to implement much narrower links.

Concurrency
Concurrency is the ability of the I/O bus to transmit data in both directions at the same time - full-duplex operation. HyperTransport I/O links support full-duplex operation. Systems based on current I/O technologies like PCI suffer arbitration issues and are not capable of delivering the bandwidth and concurrency needed for professional level DV and audio applications. When time-critical data is of the utmost importance, support for concurrency is vital.

Full Support for PCI
HyperTransport technology provides high speeds while maintaining full software and operating system compatibility with the PCI bus. The technology has been designed for concurrent connections, simultaneously running PCI cycles and other types of I/O cycles. It is designed to interface with today's I/O standards including AGP, PCI, PCI-X, IEEE-1394, USB 2.0, PL-3, SPI-4.2, and Gigabit Ethernet as well as next generation buses including AGP-8X, InfiniBand architecture, PCI-X 2.0, PCI Express, SPI-5, and 10 Gigabit Ethernet among others.

Low Latency
HyperTransport technology provides low latency access into main memory and high bandwidth to I/Os. Because of the high speeds and low latency of the channel, chips can be daisy-chained without significant performance impacts. Data streams that are latency- and bandwidth-critical receive isochronous data support to ensure ultra fast access and bandwidth to main memory.

Isochronous Support
HyperTransport technology includes support for time-dependent isochronous data such as streaming DV and real-time voice, characterized as a stream of data with packets scheduled at

regular intervals. Ensuring that synchronous data reaches its destination is critical, especially when DV is being encoded and compressed in real-time, and when the video is of movie quality.

To maximize throughput of DV and audio, the destination must receive its data with minimal delay, requiring the need for guaranteed isochronous latency and dedicated bandwidth. Without these guarantees, latency can occur, ultimately requiring more bus bandwidth. The result is sub-par frame rates, audio sync problems, and inadequate playback quality.

To support high-priority isochronous communication, the technology includes support for an operating mode in which the number of virtual channels is doubled, and associated flow control buffer types are doubled, while transactions also have an isochronous bit associated with them. In Isochronous mode there are two classes of service with a high-priority class intended to support isochronous traffic, and a low-priority class for all other traffic. High-priority is serviced first, and is prioritized in a queue so that low-priority traffic is not gridlocked while the priority stream is not taking advantage of 100% of the bus. The overall available bandwidth of the technology should be able to handle both isochronous and low-priority traffic, eliminating the need for a "fairness" algorithm. Isochronous flow control is enabled on a per-link basis to allow isochronous requests and responses to "tunnel" through non-isochronous devices on a chain.

Of course overall bandwidth within the system is still limited by the maximum and typical data rates the hard disk can provide. A HyperTransport I/O link capable of providing a data throughput of 12.8 Gbyte/s would have problems actually communicating with slower devices at this speed because the other parts within the system could not fully use the bandwidth of the bus. But a bus with this type of capability also creates opportunity, such as being able to add multiple independent links, ensuring each connection operates at maximal speed. For example, multiple PCI-X slots can be implemented so the bus is not shared. Also, extra bandwidth can be used when dealing with large data types concurrently.

NVIDIA's StreamThru technology is an excellent example of the capabilities that the technology provides. StreamThru is NVIDIA's isochronous data transport system, providing uninterrupted data streaming for networking and broadband communications. By interfacing the integrated 10/100Base-T Ethernet controller to an isochronous-aware internal bus and single-step arbiter, StreamThru assists in making streaming video and audio smoother and jitter-free.

New HyperTransport Features

Work on the HyperTransport specification continues to add features that will help in areas of DV and audio. New extensions have been included, adding peer-to-peer communications where packets can be sent directly between peer devices without having to be reflected via the host device, and 64-bit addressing that supports large memory models in excess of 1 Tbyte. Adding these extensions also includes a message-passing protocol that allows streaming a sequence of packets to a given address, and the addition of 16 streaming point-to-point flow-controlled virtual channels that support millions of end-to-end flow controlled individual streams. It also creates the ability to bridge SPI 4.2 traffic, that is typically used in communications data plane ICs, and an enhanced error recovery protocol that detects and recovers from data errors that are likely to occur when links become even faster in the future.

Multiprocessor Support For AMD Opteron
HyperTransport I/O technology is more than a system bus, it is a processor bus as well. AMD designs based on the AMD Opteron processor, with the bus technology, offer a great deal of performance for DV and audio applications. Unlike other typical configurations, these AMD multiprocessor-based systems do not share a single bus to system memory with other devices within the system. Depending on how they are arranged, an AMD multiprocessor-based system using two, four, or even eight processors, can seamlessly communicate amongst themselves using three links that are built directly into the processor die. And since the technology offers bandwidth to spare, the latencies from processor-to-memory are significantly low while performance yields will continue to increase as advances are made in memory technology.

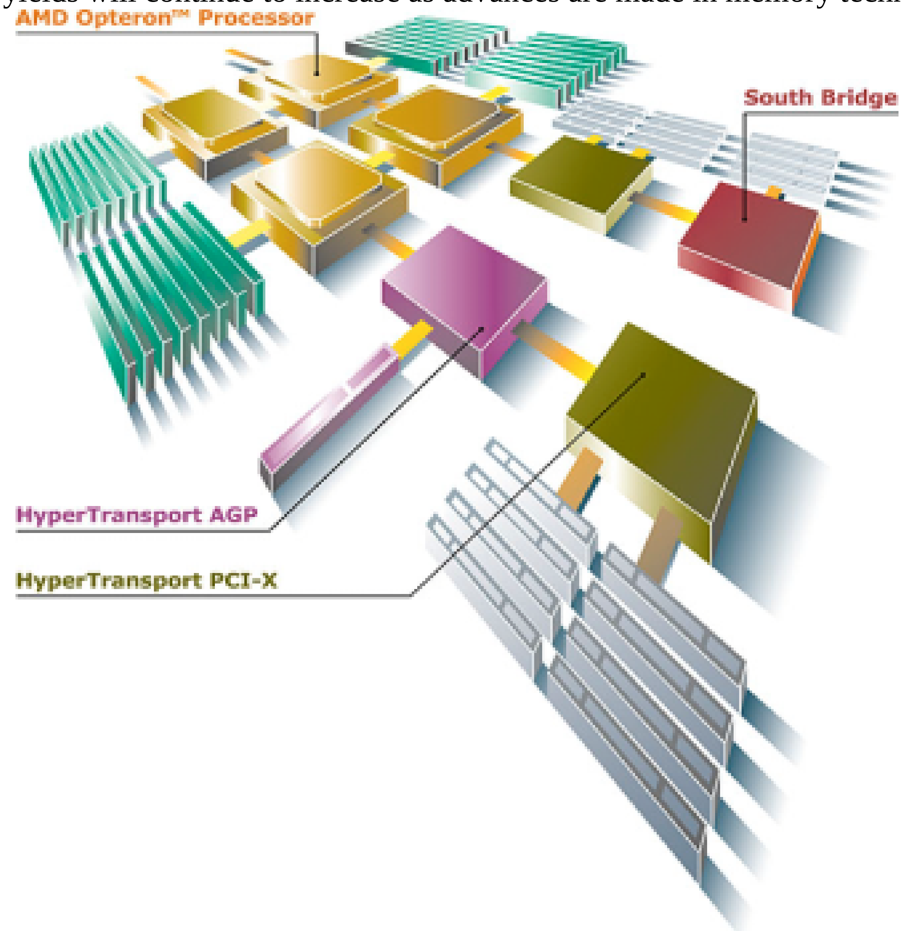

Fig. 3: Quad-Processor Platform Based on the AMD Opteron

Integrated DDR DRAM Memory Controller
Memory bandwidth between the system memory and the processor core has been one of the greatest limiting factors in performance with regards to the real-time editing of high-resolution video and audio. The AMD Opteron Athlon processors, based on Hammer technology, directly address this bottleneck by integrating a memory controller into the processor, completely changing and revolutionizing the method for the way x86-based processors access main memory. By attaching memory directly to the processor, moving the memory controller from the

Northbridge to reside directly on the processor, and eliminating the front-side bus altogether, the processor can benefit from increased memory bandwidth with an overall reduced latency.

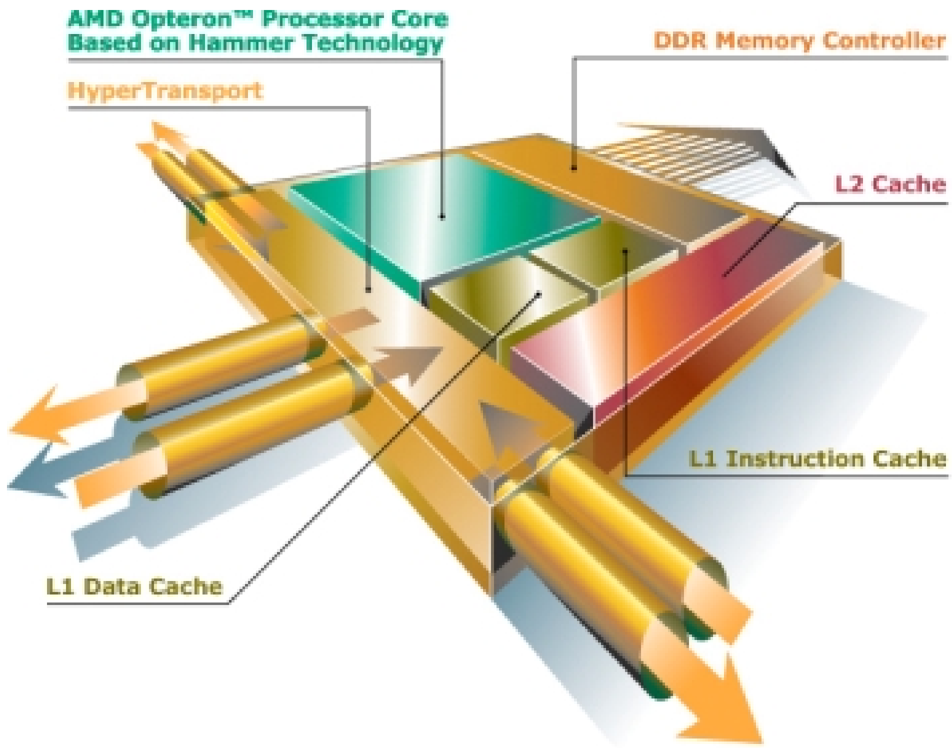

Fig. 4: Functional Diagram of an AMD Processor Based on Hammer Technology

AMD processors based on Hammer technology may incorporate a dual-channel DDR DRAM controller with a 128-bit interface capable of supporting up to eight DDR DIMMs (four per channel.) With PC2700 memory, rated at speeds of 333 MHz, the available memory bandwidth available to the processor becomes equivalently 5.3 Gbyte/s. Since the memory controller is now operating at the same GHz speeds as the processor latency is further reduced.

But it doesn't stop there. The integrated memory controller delivers even more scalability in multiprocessor designs. Taking the example above, with PC2700 memory, but this time within a four-processor multiprocessing system based on the AMD Opteron and support for up to 32 DIMMs, the overall memory bandwidth is designed to deliver an astonishing 21.3 Gbyte/s of available memory bandwidth, exceeding the requirements for the largest DV and audio streams.

Summary

HyperTransport technology, thanks to its high-bandwidth, low latency connections, isochronous transport system, and support for concurrent communications, offers the bandwidth and performance necessary for the next generation of DV and audio. The technology provides backwards compatibility for PCI software, drivers, and OSs, while helping eliminate bottlenecks and providing the bandwidth necessary for future high-speed chips and interconnects standards.

The is licensed royalty-free to all members of the HyperTransport Technology Consortium. More information can be found at the consortium's web site www.hypertransport.org